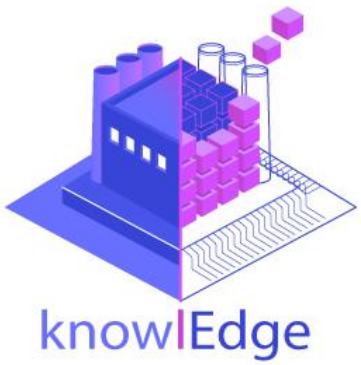
Towards AI powered manufacturing services, processes, and products in an edge-to-cloud-knowlEdge continuum for humans [in-the-loop]



WP3: Data Integration, Governance, and Quality Management

EU ID: D3.3 Initial Data Management and Data Quality modules v1.0

Deliverable Lead and Editor: Xu Tao, LINKS Foundation

Contributing Partners: T3.2 partners

Date: 2021-12

Dissemination: Public

Status: For EU Approval

Abstract

The purpose of this knowlEdge deliverable, is presenting the initial work on the Data Management and Data quality Management module.

Grant Agreement: 957331

Document Status

Deliverable Lead	Xu Tao, LINKS Foundation
Internal Reviewer 1	Nicolò Bertozzi, LINKS Foundation
Internal Reviewer 2	Rosaria Rossini, LINKS Foundation
Internal Reviewer 3	Christian Beecks, WWU
Internal Reviewer 4	Stefan Walter, VTT
Type	Deliverable
Work Package	WP3: Data Integration, Governance, and Quality Management
ID	D3.3 Initial Data Management and Data Quality modules v: 1.0
Due Date (Original)	2021-12
Delivery Date	2021-12
Status	V.1.0

History

See Annex B.

Status

This deliverable is subject to final acceptance by the European Commission.

Further Information

www.knowlEdge-project.eu and mailto:info@knowlEdge-project.eu

Disclaimer

Project Partners:

For full details of partners go to www.knowlEdge-project.eu/partners

Executive Summary

The purpose of this knowlEdge deliverable, D3.3, is to give the initial vision of the design and implement the Data Quality Assurance Framework which will be deployed in the edge of the overall knowlEdge Architecture. It plays the key role to guarantee the data quality before applying the data analysis and machine learning methods. KnowlEdge aims to develop solutions for the industrial plants for improving the performance, efficiency, economic. Therefore, assuring the data quality is crucial to make more accurate decisions for the industrial operations.

In the first section an introduction on knowlEdge project is shown. The second section will introduce the definition of the Data Refinement and Quality with its current Background and the existing challenges.

Table of Contents

1 Introduction

0.1 knowlEdge Project Overview

The knowlEdge project is funded by the H2020 Framework Programme of the European Commission under Grant Agreement 957331 and conducted from January 2021 until December 2023. The knowlEdge consortium consists of 12 partners from 7 EU countries, and its solution will be tested and evaluated in 3 manufacturing sectors with a total budget of circa 6M€. Further information can be found at www.knowlEdge-project.eu

AI is one of the biggest mega-trends towards the 4th industrial revolution. While these technologies promise business sustainability and product/process quality, it seems that the ever-changing market demands and the lack of skilled humans, in combination with the complexity of technologies, raise an urgent need for new suggestions. Suggestions that will be agile, reusable, distributed, scalable, accountable, secure, standardized and collaborative.

To break the entry barriers for these technologies and unleash their potential, the knowlEdge project will develop a new generation of AI methods, systems and data management infrastructure. This framework will provide means for the secure management of distributed data and the computational infrastructure to execute the needed analytic algorithms and redistribute the knowledge towards a knowledge exchange society. To do so, knowlEdge proposes 6 major innovations in the areas of data management, data analytics and knowledge management: (i) A set of AI services that allow the usage of edge deployments as computational and live data infrastructure, an edge continuous learning execution pipeline; (ii) A digital twin of the shop-floor to test the AI models; (iii) A data management framework deployed from the edge to the cloud ensuring data quality, privacy and confidentiality, building a data safe fog continuum; (iv) Human-AI Collaboration and Domain Knowledge Fusion tools for domain experts to inject their experience into the system to trigger an automatic discovery of knowledge that allows the system to adapt automatically to system changes; (v) A set of standardization mechanisms for the exchange of trained AI-models from one context to another; (vi) A knowledge marketplace platform to distribute and interchange AI trained models.

0.2 Component Purpose and Scope

In this document, the initial vision of the design and implementation of the Data Quality Assurance Framework is presented. The component will be deployed in the edge of the overall knowlEdge Architecture and plays a key role for the data quality guarantee prior to the application of data analysis and machine learning methods. KnowlEdge aims to develop solutions for industrial plants in order to improve their performance, efficiency, and economics. Therefore, assuring the data quality is crucial to make more accurate decisions for the industrial operations.

The corresponding component and its position within the initial knowlEdge architecture is marked in a red square in Figure 1.

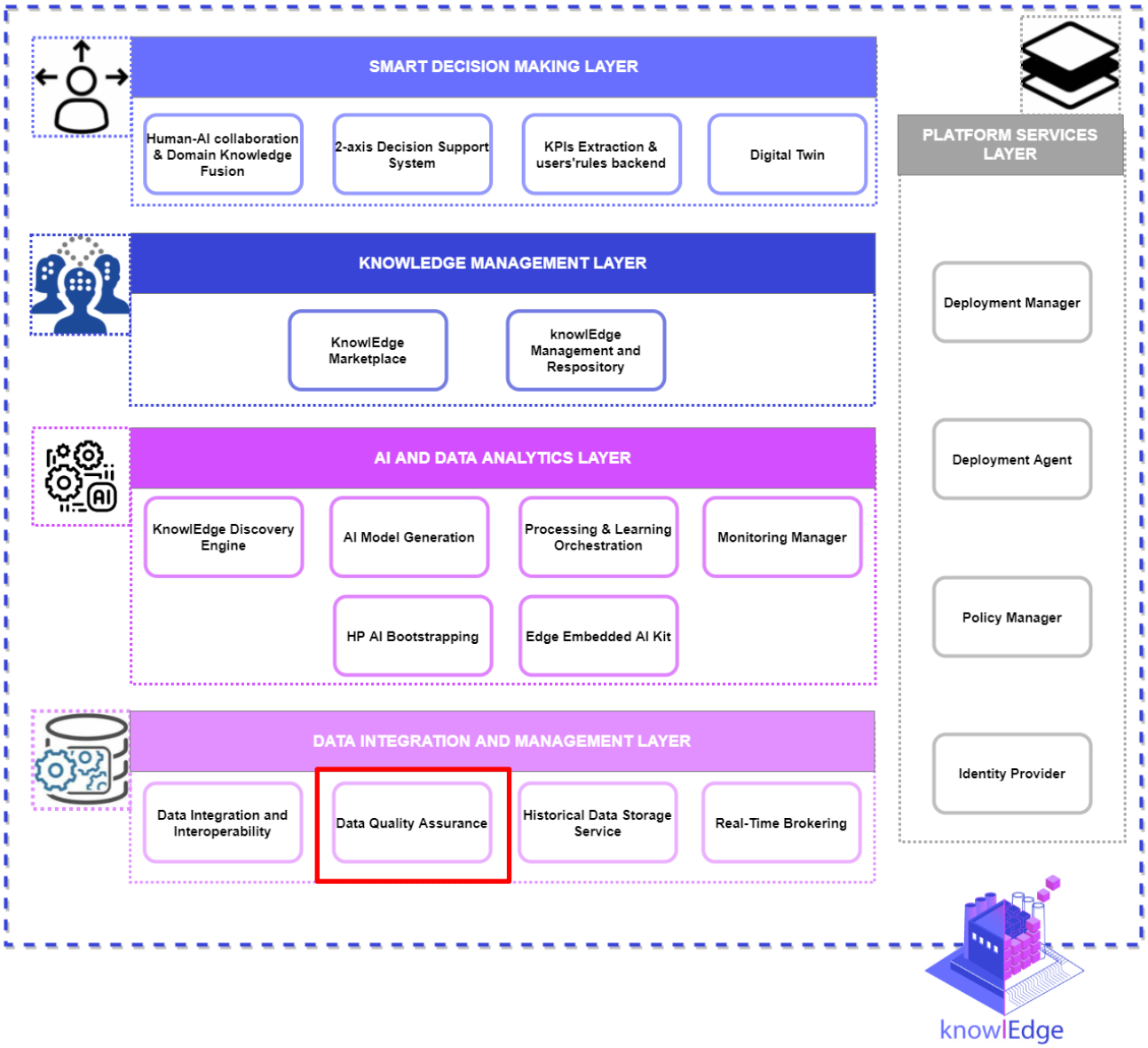


Figure 1. knowlEdge Global architecture

0.3 Deliverable Context

This manual is based on the project procedures as defined within the knowlEdge Description of Action and Consortium Agreement and extends them in the operational aspects where necessary. However, it is subservient to those documents.

0.4 Document Status

This document is listed in the Description of Action as “confidential” since it provides information for project-internal usage only.

0.5 Document Dependencies

This document has no preceding documents or expected further formal iterations. If explicitly requested by reviewers, a definitive version can be made available at the end of the project.

0.6 Glossary and Abbreviations

A definition of common terms related to knowledge, as well as a list of abbreviations, is available at www.knowlEdge-project.eu/glossary

0.7 External Annexes and Supporting Documents

External Documents:

- Annexes:
  - none
- Supporting Documents:
  - none

0.8 Reading Notes

- None

0.9 Document Updates

None

1 Background, Challenges and Requirements of Data Refinement and Quality

In this section, we will introduce the definition of the Data Refinement and Quality with its current Background and the existing challenges. Then, the requirements and expectations from the knowlEdge project will be further illustrated.

1.1 Background

With the rapid technology development of IoT (Internet of Things) and Industry 4.0, the number of the connected devices is increasing significantly. And technologies such as the WWW(World Wide Web), engineering and science applications and networks, business services and many more are generating data in exponential growth due to the development of powerful storage and connection tools. As a consequence, the scale of data to be stored, managed, and processed showed a massive growth. The tremendous data facilitates employment of Big Data and Machine Learning solutions in the ICT(Information and Communication Technology) sectors to help the enterprises increasing their commercial values. It has been applied massively across all the society sectors, such as healthcare, transportation, manufacture factory, agriculture and e-commercials. These new technologies, products, systems, and services are expected to create an annual economic impact of $2.7 trillion to $6.2 trillion by 2025 [1].

However, the data generated are frequently characterized by high volume, high velocity and high variety which require high-performance processing [2]. Nevertheless, the vast amounts of raw data cannot be directly used by humans or applications to obtain the well-understood knowledge and information. This paradigm resulted in the research and development of emerging data science discipline, which is playing a significant role in the current information age. To maximize the benefits of the emerging Big Data and Machine Learning technologies, ensuring a high degree of data quality is crucial and a fundamental concern in the design of IoT based products and services. For example, in March 2019, Tesla's Autopilot was engaged in a fatal crash of a Tesla electric vehicle, because the data coming from the vehicle's self-driving sensor (i.e. radars) did not match with actual road situations, failing to detect objects crossing the road and causing the vehicle to crash into a truck [3]. To date a growing body of research studies have investigated data quality (DQ) focusing on aspects such as: DQ dimensions, DQ problems, and techniques to improve DQ in IoT.

KnowlEdge aims to provide innovative solutions in the areas of data management, data analytics and knowledge management. One of the relevant objectives is developing a data management framework from the edge to the cloud ensuring data quality, privacy and confidentiality, building a data secure fog continuum. The work in this deliverable addresses the needs for data quality assessment and data quality improvement with the consideration of the personal data privacy and confidentiality within the overall knowledge Data Management Framework. This work will be considered as an important cornerstone to improve the performance and accuracy of the AI (Artificial Intelligence) solutions proposed within knowledge project.

The next subsection analyses the existing challenges concerning the Data Refinement and Quality Assurance.

1.2 Challenges

The characters of data in IoT come down to the 4Vs: Volume, Velocity, Variety, and Value. Extracting high-quality and real data from massive, variable, and complicated data sets becomes an urgent issue. The main challenges are as following:

C1: The diversity of data sources brings abundant data types and complex data structures and increases the difficulty of data integration.

The data is generated from a wide range of sources including: 1) data sets from the internet and mobile internet, 2) data from the Internet of Things, 3) data collected by various industries, and 4) scientific experimental and observational data, such as high-energy physics experimental data, biological data, and space observation data. In addition, one data type is unstructured data, for example, documents, video, audio, etc. The second type is semi-structured data, including: software packages/modules, spreadsheets, and financial reports.

C2: Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.

It is difficult to collect, clean, integrate and finally obtain the necessary high-quality data within a reasonable time frame.

C3: Data changes very fast and the "timeliness" of data is very short, which necessitates higher requirements for processing technology.

Due to the rapid changes in big data, the "timeliness" of some data is very short. If the required data cannot be collected in real time or the time of dealing with the data needs over a very long time, then they may obtain outdated and invalid information.

C4: No unified and approved data quality standards have been formed, and research on the data quality of big data just begun.

There are many disputes about it, some standards need to be mature and perfectioned. In fact, quality applied to data has various definition and the scenario drives the needs of what standard has to be applied.

a. Requirements derived within knowlEdge Project

Based on the challenges illustrated above, the data quality assessment will mainly focus to cope with the challenges forementioned, in this deliverable. And some requirements within the knowlEdge project scope will be defined in the second period (update/final document) in close collaboration with the pilots and tailored for each scenario.

Challenge	Aim

Figure 2: table

2 Methodology of Data Assurance Framework

b. Data Quality Criteria

Data quality depends not only on its own features but also on the business environment which uses the data. There is no unified standard to define the criteria of high data quality. Usually, data quality standards are developed from the perspective of data producers. Nowadays, with the diversity of data sources, data users are not necessarily data producers. Thus, it is very difficult to measure data quality. The commonly used indicators for data quality measurement [4] are shown in Figure 3.

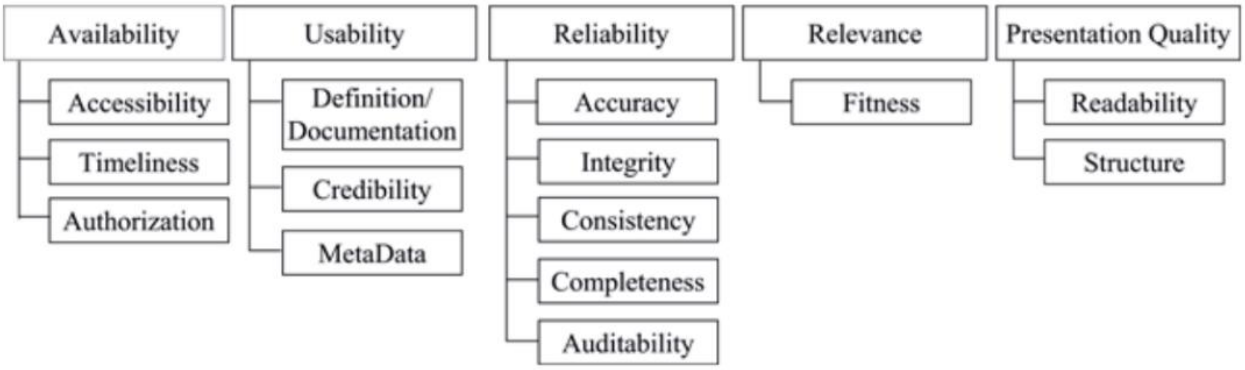


Figure 3: Data Quality Criteria

c. Causes of imperfect data

Based on the criteria listed in the Figure 3, there are some possible reasons that could violate the data quality, as listed below:

Causes of non-accuracy

1) wrong placement or selection of sensors

2) outliers due the sensor faults

3) noise of undesired change that deviates from the original signal, such as exhausted batteries, faulty memory cells, bit error in transmission or interference when multiple wireless devices transmit the data simultaneously on the same frequency bands.

Causes of non-Timeliness

Missing updates and low data rate, such as in the context of agriculture, devices have constrained resources such as energy and are required to communicate across large distances using technologies such as LoRaWAN or SigFox, which are prone to low data rates and high latency but require very little energy.

Cause of non-Completeness

It refers to data availability and missing data, missing data can be caused by sensor inefficiencies, communication issues or by attacker's intercepting or manipulating data, the

lack of data updates. In addition, the data owners selectively disclose the data based on certain constraints (e.g. privacy considerations), resulting in less detailed data being available for users.

Cause of non-credibility

Noise caused by the IoT devices transmitting data simultaneously. The data loss or missing data could decrease the utility of IoT data.

Data volume issue

It might be due to the transmission rate, when there is large amount of data to transmit, for example, when IoT devices collect video and image data that requires data compression and subsequent sound or image recovery, data loss and delay are also accompanied in this process.

d. Methods for the Data Quality measurement

There are a set of technologies for measuring the DQ as following:

1) **Measurement between techniques, sources or defined attributes (MTS):** Data elements in two or more IoT datasets that are derived from different techniques, sources or defined attributes are compared to see if there are agreements in these elements. For example, the IoT datasets can be collected by different experimental settings (e.g. take distances between a transmitter and a receiver into account, protocols, data sources) or algorithms.

2) **Measurement with a reference (MR):** A dataset derived from another source serves as a reference to compare with the collected IoT dataset to determine whether or not there are agreements in these elements. This method can be used to measure the completeness and accuracy by referring to the results with prior literature that used the same IoT dataset for actual values, an applicable range of values, historical data, or spatial-temporal correlated measured values for the spatial-temporal correlated measured values for the objects provided by the sensor and its neighbours as a reference for DQ measurement.

3) **Devices or algorithms validation (DAV):** The collected IoT dataset is examined by using well developed devices or algorithms to ascertain whether or not expected values are present. For example, the dataset can be divided into a training dataset and a testing dataset and then measured the accuracy by looking at the agreement between the results of the testing dataset and the expected values using the proposed approaches implemented on the training dataset.

4) **Measurement between time intervals (MTI):** The IoT dataset is examined during a fixed time interval to determine how good is the data collected. This method can be used to divide an IoT dataset into data slices based on a certain temporal duration. And observe how much data was accessed by consumers from the dataset over the time for the availability.

5) **Measurement of presence (MP):** The collected IoT dataset is examined to determine whether or not data elements are present and thus describe the completeness of the IoT dataset.

6) **Process observation (PO):** The loading process and physical association of sensors are monitored to ascertain whether or not the data collected make sense.

7) **Log files review (LR):** Log files and/or claims investigations on the collected IoT data are reviewed to determine whether or not data errors or anomalies present.

e. Methodology of Data Quality Assurance Framework

The data acquisition is relatively easy, but since most of the data collected does not always satisfy internal quality requirements, we need to improve data quality as far as possible under these conditions without a significant increase in acquisition cost. Usually there are data problems such as data errors, missing information, inconsistencies, noise, etc. The flow shown in Figure 4 can be used as a guideline [4] to define the data quality assurance framework to guarantee the data quality under the defined dimensions with indicators.

Link between data quality dimensions, manifestations of DQ problems, and methods utilized to measure data quality. The data quality mainly focuses on aspects such as: DQ dimensions, DQ problems, and techniques to improve DQ in IoT. The quality has been defined both as “fitness for use” and as “conformance to requirements”. There are two important concepts used in the study are dimension and manifestation. A dimension of DQ refers to an individual aspect of DQ such as completeness and accuracy. Manifestation of DQ problems is defined as a symptom of, or challenge arising from, data errors or anomalies.

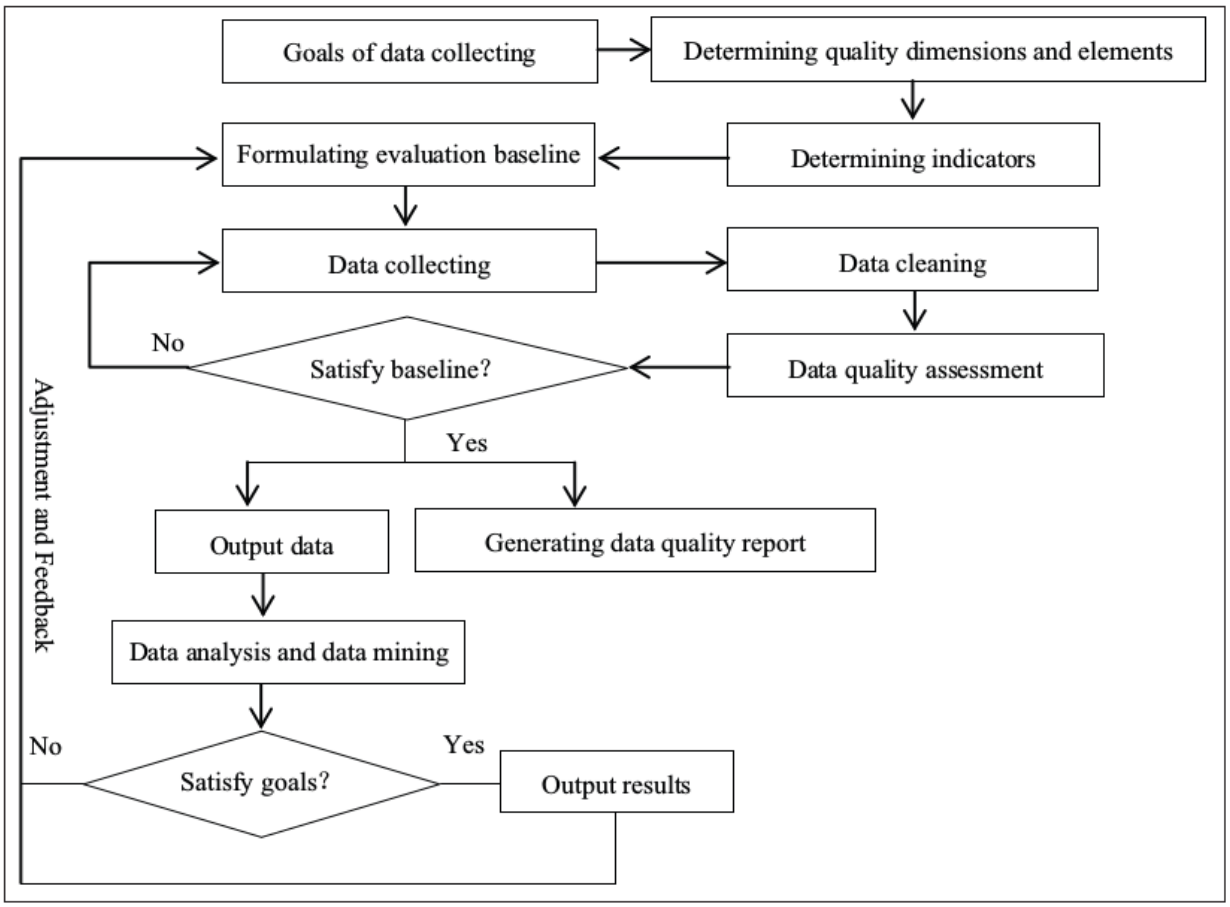


Figure 4: Data Quality Assurance logic framework

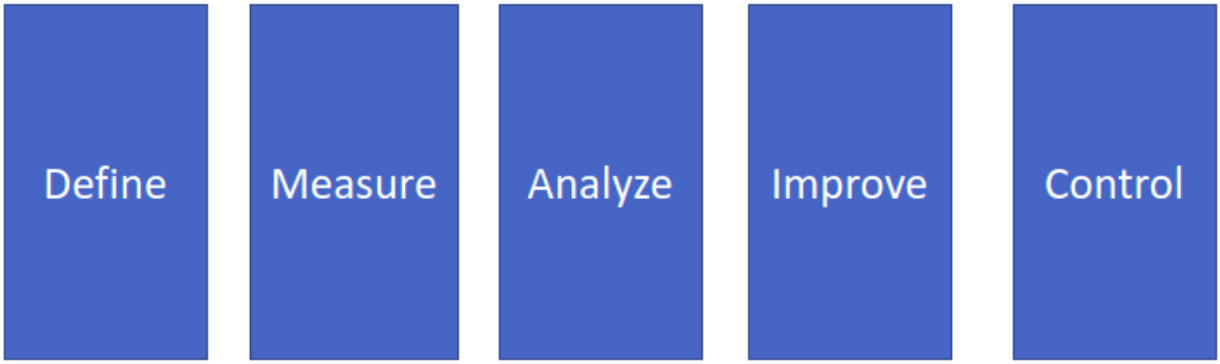


Figure 5: Methodology of Initial Data Quality Management Framework

The initial Data Quality Assurance framework is proposed based on the logic showed in Figure 4, and the refined Six Sigma [5] strategies in Figure 5, which aims to improve manufacturing quality by identifying and removing the causes of defects and minimizing variability in manufacturing and business processes. In KnowlEdge, we refined the strategy to adapt the needs of the data quality assurance.

The approach to design and develop a data quality assessment framework is:

1) **Definition of the DQ Requirements and a set of dimensions**, such as Accuracy, Timeliness, Completeness, Timeliness, and Utility as listed in the figure 1 above.

2) **Measurement of the data quality,** Mapping the defined requirements and quality dimensions to a set of metrices, statistical formulars for validating the data quality. The technics could be Measurement among techniques, sources or defined attributes; Measurement with a reference; Devices or algorithms validation; Measurement within time intervals; Measurement of presence; process Observation; Log files review.

3) **Analysis to identify the manifestations of the DQ problems,** these manifestations are: measurement errors, noise, artifact error, data frame distortion, dirty data, outliers, missing data, missing updates, data loss, and delay data transmission.

4)**Design and development solutions** to address and/or improve IoT data quality by variety of solutions. These solutions could include: 1) protocols for data transmission 2) frameworks for storing IoT data, collecting sensor data, and monitoring the delivered IoT data. 3) architectures for monitoring DQ and filtering good data from collected IoT data, cleaning IoT data streams, and providing data products. 4)tools for updating real-time data to cloud, identifying data anomalies and dealing with missing data to address and/or improve DQ in IoT.

4) **Improve:** Implementation of quality improvement methods and solutions to address and/or improve IoT data quality by variety of solutions. These solutions could include: 1) protocols for data transmission 2) frameworks for storing IoT data, collecting sensor data, and monitoring the delivered IoT data. 3) filtering good data from collected IoT data, cleaning IoT data streams, and providing data products. 4)tools for updating real-time data.

5) **Control:** Monitoring the data quality periodically and/or reporting data quality issue.

3 KnowlEdge Data Quality Assurance Framework

This section will present the initial architecture of Data Quality Assurance Framework defined by following the methodology in section e. In KnowlEdge, the project partners intend to provide a comprehensive framework for data management to enable data collection, data quality assessment, data quality improvement, and data monitoring. The initial Data Quality Assurance Framework is shown in Figure 6.

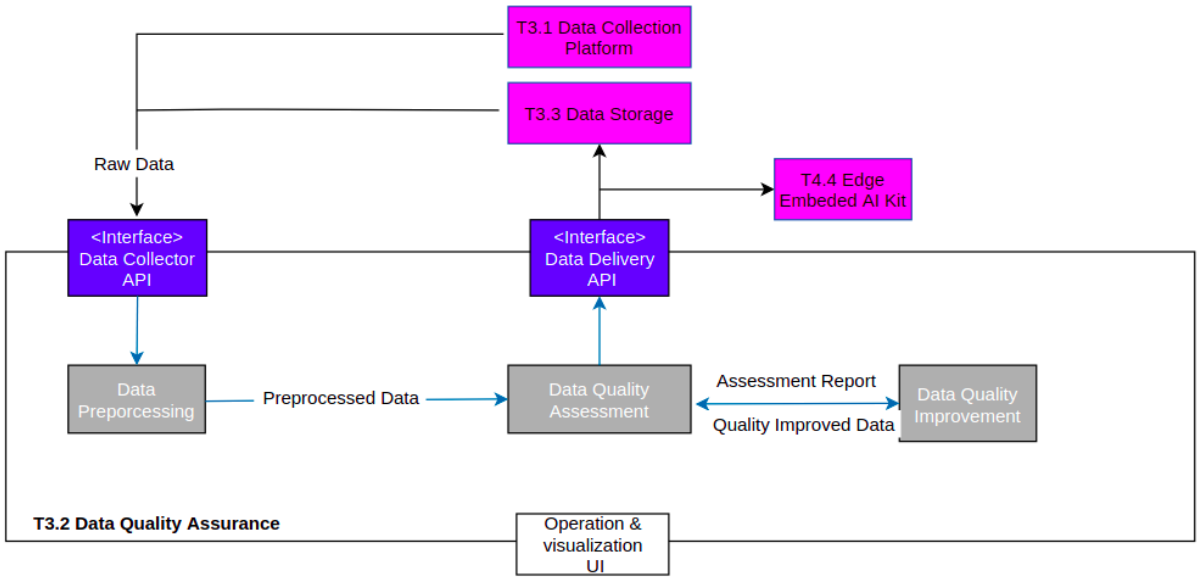


Figure 6: Initial Data Quality Management Architecture

As can be seen in the figure above, the framework is composed of four main components in the initial architecture: Operation & Visualization UI, Data Preprocessing, Data Quality Assessment, Data Quality Improvement.

- **Operation & Virtualization UI** provides the user interface to the data scientists to define and manage the data requirements and dimension rules for data quality assessment. Furthermore, it allows us to generate data assessment reports and to visualize data in an intuitive way.
- **Data Preprocessing** performs the preprocessing operations to the raw dataset such as filtering, eliminating duplicates, and formatting, etc.
- **Data Quality Assessment** validates and measures data quality using the metrics and algorithms developed from the defined data requirements and dimension rules.
- **Data Quality Improvement** provides solutions to improve the data quality according to the data quality assessment report.

3.1 Technical Foundations

After walking though some state-of-the-art technologies, we identified the initial technologies to be used for the implementing the architecture.

The data to be processed will be taken from the KnowlEdge data storage or from the Data collection platform. Then these raw data will pass through the Data Preprocessing implemented by Python and Pandas[1] for some simple operations. After that, the data quality will be measured with Data Quality Assessment which will be implemented with a set of measurement algorithms, the assessment report will be displayed in the Operation & Virtualization UI. Based on the assessment results, corresponding data quality improvement operations will be performed leveraging the solutions implemented by Python and Pandas. Afterwards, the Improved dataset will be validated another time to check whether the requirements are satisfied or not. Finally, the improved dataset will be either stored in the Knowledge database or be delivered to Edge Embeded AI kit through the defined API. The Operation & Visualization UI can be implemented by Thingsboard[2], HTML, and Flask[3].

1 https://pandas.pydata.org/
2 https://thingsboard.io/
3 https://flask.palletsprojects.com/en/2.0.x/

4 Conclusion

In this document the initial work on Data Management and quality has been proposed. The initial architecture is presented according to the knowledge platform.

As a starting point of this task, a study on data quality technic has been conducted. In this study the general behavior of the component and then create the architecture accordingly.

Furthermore, possible technologies have been identified in other to provide suitable solutions for the individual components.

References

[1] M. C. J. B. R. D. P. B. A. M. James Manyika, “Disruptive technologies: Advances that will transform life, business, and the global economy,” 2013.

[2] D. Laney, “3D Data Management: Controlling Data Volume, Velocity, and Variety,” 2001.

[3] A. Villasanta, International Business Times, 2019. [Online]. Available: https://www.ibtimes.com/tesla-model-3-autopilot-feature-blame-death-driver-crash-2792690.

[4] Y. Z. Li Cai, “The Challenges of Data Quality and Data Quality,” *Big Data Era. Data Science Journal,* pp. 1-10, 2015.

[5] [Online]. Available: https://en.wikipedia.org/wiki/Six_Sigma.

Annex A: History

Document History	
Versions	• 01 Initial version with structure • 0.9 Draft for internal discussion • 1.0 Final version
Contributions	• Xu Tao (LINKS)